CSI5126. Algorithms in bioinformatics

Overview of the course **content** and **expectations**

Marcel Turcotte

uOttawa

School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

Version September 6, 2018

Motd: scholarships

- www.uottawa.ca/graduate-studies/students/awards
- Intellectual independence
- Building up your curriculum vitæ
- Natural Sciences and Engineering Research Council of Canada
- NSERC, CIHR, OGS, ...

Introduction

I want to learn about **you**.

- What is your **name**?
- Are you an **undergraduate** or a **graduate** student?
- If you are a **graduate** student:
  - Who is your **supervisor**?
  - Give us two or three sentences about your research topic.
- Where are you from?
- What background do you have in **biology**?
- Do you program in **Java** (at least the basics)?
- What are you **expecting** from this course?

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules

- 1989, **Honours project**, implementation of a graphical user interface for a protein folding/unfolding system
- 1989–95, **Université de Montréal**, graduate studies under the direction of Guy Lapalme (IRO), Robert Cedergren (Biochemistry), work on methods for building nucleic acids' 3-D structures
- 1995–97, **University of Florida**, work with Steven A. Benner (Chemistry) on evolutionary-based approaches to predict protein secondary structure
- 1997–00, **Imperial Cancer Research Fund** (London/UK), work with Michael J.E. Sternberg and Stephen H. Muggleton (York) on the application of Inductive Logic Programming to discover automatically protein folding rules
- 2000–, **University of Ottawa**, work on nucleic acids secondary structure determination, motifs inference and pattern matching

What is **bioinformatics**?

TED: Juan Enriquez on genomics and our future

http://www.ted.com/talks/juan_enriquez_on_genomics_and_our_future.html

A. Isaev

"Broadly speaking, bioinformatics can be defined as a collection of **mathematical, statistical and computational methods for analyzing biological sequences**, that is, DNA, RNA and amino acid (protein) sequences."

In *Introduction to Mathematical Methods in Bioinformatics*,
A. Isaev, Springer, p. i, 2006.

Lacroix and Critchlow

"Bioinformatics is the design and development of computer-based technology that supports life sciences. Using this definition bioinformatics tools and systems perform a diverse range of functions including: **data collection**, **data mining**, **data analysis**, **data management**, **data integration**, **simulation**, **statistics**, and **visualization**. *Computer-aided technology directly supporting medical applications is excluded from this definition and is referred to as medical informatics*."

In *Bioinformatics: Managing Scientific Data*, **Zoé Lacroix and T. Critchlow** Editors, Morgan Kaufmann, p. 3, 2003.

Jones N.C. and Pevzner P. A.

"Biologists that reduce bioinformatics to **simply the application of computers in biology** sometimes fail to recognize the rich intellectual content of bioinformatics. Bioinformatics has become a part of modern biology and often dictates new fashions, **enables new approaches**, and **drives further biological developments**"' In *An Introduction to Bioinformatics Algorithms*, **Jones N.C. and Pevzner P. A.**, MIT Press, p. 77, 2004.

J.J. Ramsden

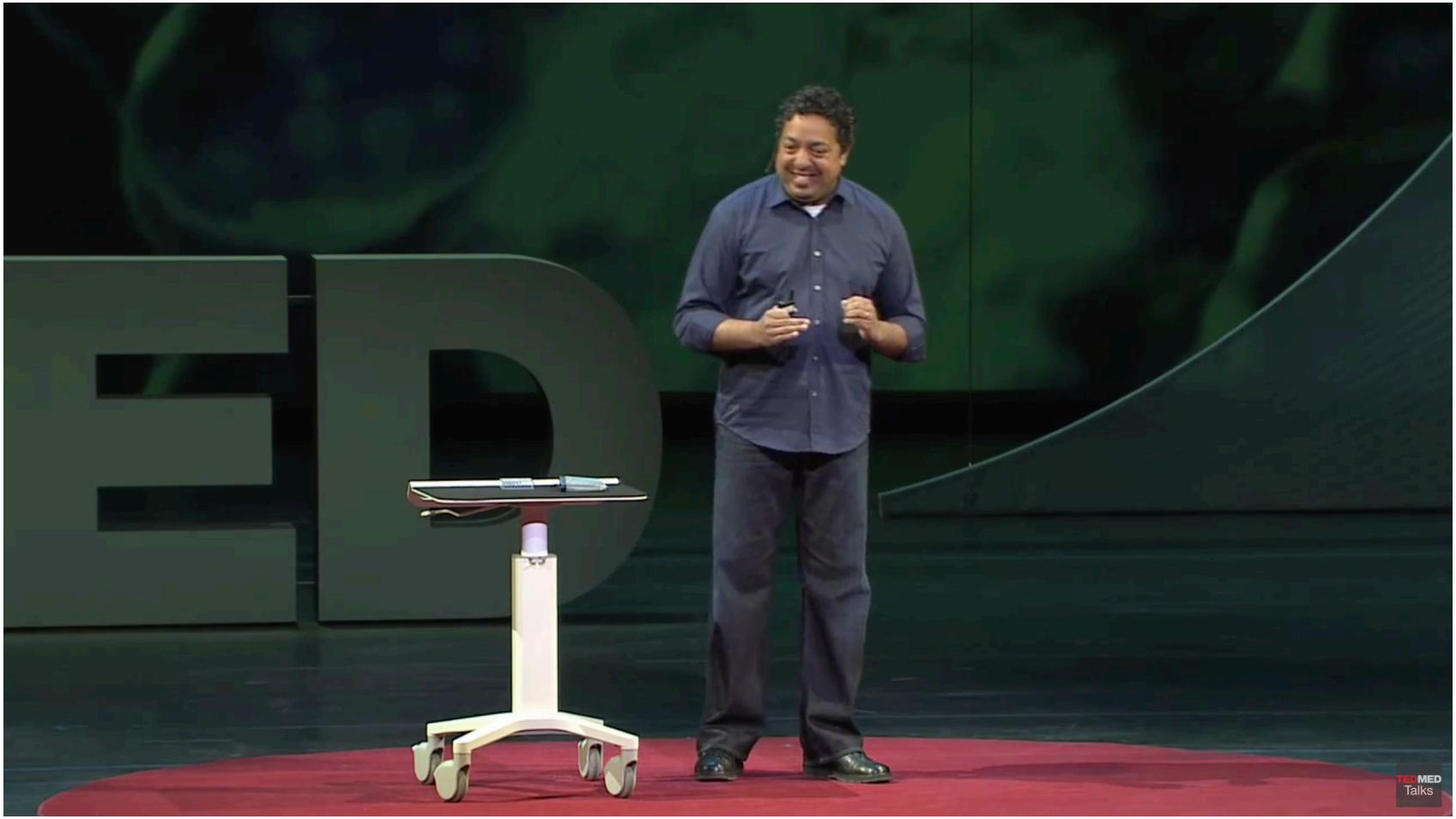
"In bioinformatics, so much is to be done, the raw material to hand is already so vast and vastly increasing, and the problems to be solved are so important (perhaps the most important of any science at present) **we may be entering an era comparable to the great flowering of quantum mechanics in the first three decades of the twentieth century** (...)"

In *Bioinformatics: An introduction*, **J.J Ramsden**, Kluwer, p. xiii, 2004.

SIB - Swiss Institute of Bioinformatics

https://youtu.be/182AzhLiwxc

Atul Butte/Stanford at TEDMED 2012



https://youtu.be/dtNMA46YgX4

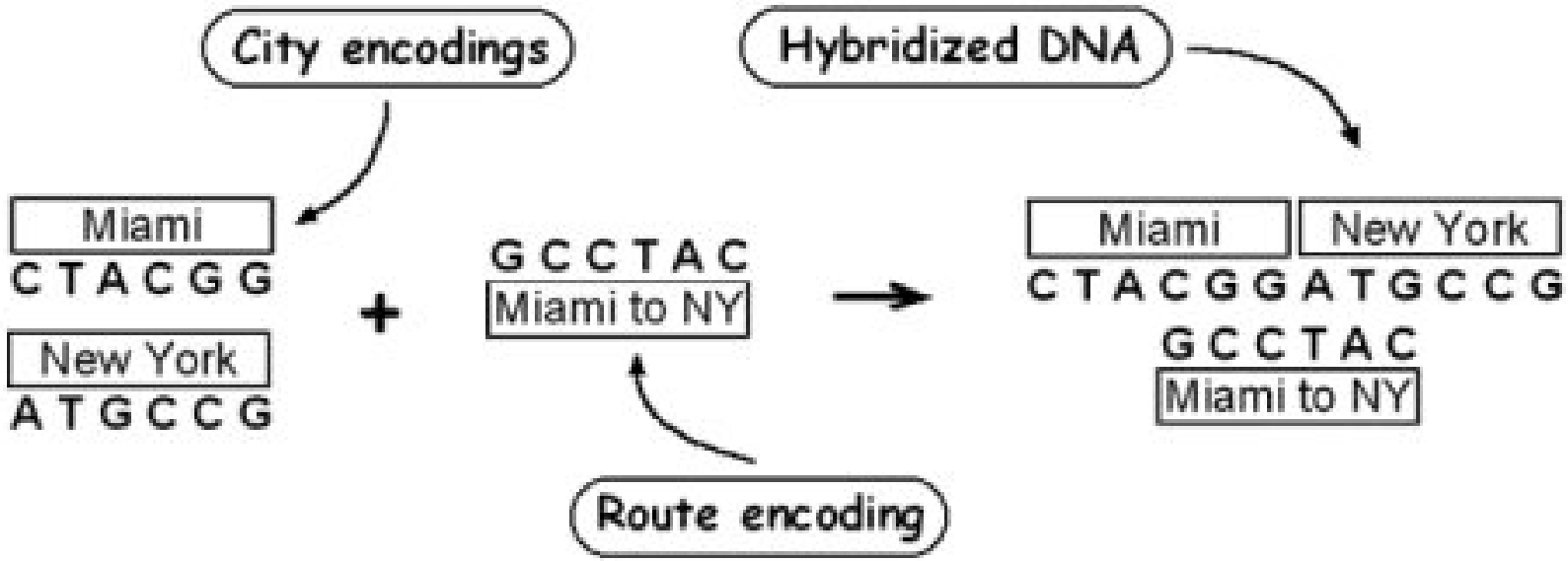
Origins

"Computers and specialized software have become an essential part of the biologist's toolkit. Either for routine DNA or protein sequence analysis or to parse meaningful information in massive gigabyte-sized biological data sets, virtually all modern research projects in biology require, to some extent, the use of computers. (...) **the very beginnings of bioinformatics occurred more than 50 years ago**, when desktop computers were still a hypothesis and DNA could not yet be sequenced."

Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief Bioinform* 79, 137 (2018).

What it's **not**!

Leonard **Adleman** (*Science*, December 1994) solved a particular instance of the Hamiltonian Path problem using DNA molecules!



$\Rightarrow$ An Hamiltonian path visits every node of a graph exactly once.

What it's **not**! (contd)

DNA computing is the theoretical study of the use of DNA molecules to solve challenging problems or as a new architecture (what class of problems can be solved, what are the properties, limits, etc.).

What it's **not**! (contd)

Biotechnology and **biomedical engineering** apply engineering approaches to problems dealing with biological systems.

Examples of biomedical engineering include **developing biomedical devices** for human implantation, **drug delivery systems**, **simulation of organs and micro-fluids**, **medical imaging**, and many more.

Other bioinformatics courses on campus

- `http://www.bioinformatics.uottawa.ca`
- **BNF5106** Bioinformatics*
- **BCH5101** Analysis of -omics data
- Ottawa Bioinformatics User Group (**OttBUG**)
  - See **reddit** conversation

*www.bioinformatics.uottawa.ca/stephane/bnf5106.syllabus.pdf

Joint collaborative program in bioinformatics (MSc)

Starting from January 2008, Carleton University and the University of Ottawa offers a Collaborative Program leading to an **MSc degree with Specialization in Bioinformatics** or **MSc of Computer Science degree with Specialization in Bioinformatics**.

Course **learning outcomes**

Upon completion of the course, student will be able to:

- **List** and **describe** the fundamental algorithms in bioinformatics
- **Articulate** the trade-offs behind algorithms in bioinformatics
- **Write** computer programs for solving large scale bioinformatics problems
- Critically **review** scientific publications in this field
- **Locate** and critically **evaluate** scientific information
- **Apply** one of the paradigms presented in class to solve real-world problems
- **Present** scientific content to a small technical audience

Outline

- Essential cell biology
- Suffix trees, lowest common ancestor
- Suffix trees applications
- Molecular sequence alignment
- Students presentation (review paper)
- Phylogeny
- RNA secondary structure
- Sequence motifs (deterministic and probabilistic)
- Students presentation (final project)

Evaluation

- Programming assignments (20%)
- Review and oral presentation of a scientific publication (10%)
- Midterm examination (20%)
- Project (50%) - (proposal 10%, presentation 10%, report 40%)

Textbooks

The following textbooks can be downloaded freely as PDF (access restricted to uOttawa IP addresses).

- Bernhard Haubold and Thomas Wiehe (2006). *Introduction to computational biology: an evolutionary approach*. Birkhäuser Basel.
- Wiesława Widłak (2013). *Molecular Biology: Not Only for Bioinformaticians* (Vol. 8248). Springer.
- Warren J. Ewens, Gregory R. Grant (2001) *Statistical Methods in Bioinformatics: An Introduction*. Springer.

Other excellent textbooks.

- Wing-Kin Sung (2010) Algorithms in Bioinformatics: A Practical Introduction. Chapman & Hall/CRC. QH 324.2 .S86 2010
- Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchinson (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press.
  - QP 620 .B576 1998
- Dan Gusfield (1997) *Algorithms on strings, trees, and sequences : computer science and computational biology*. Cambridge University Press.
  - QA 76.9 .A43 G87 1997
- Pavel A. Pevzner and Phillip Compeau (2018) Bioinformatics Algorithms: An Active Learning Approach. Active Learning Publishers. `http://bioinformaticsalgorithms.com`

Jobs

If you want to compete in bioinformatics, first you need to compete for really smart people. You need really smart people who understand how to manipulate nanomolecules.

Juan Enriquez

Jobs: http://www.bioinformatics.ca/jobs

Job Postings

DISCLAMER: OICR and the Canadian Bioinformatics Workshops are not affiliated with and have not investigated the companies listing jobs on this site. OICR is not making any representations with respect to the positions and is not acting as an agent for the companies listed. The OICR and CBW reserve the right to select and edit job postings which are added to this site to ensure Canadians are eligible to apply and that positions are bioinformatics related.

HOW TO POST A JOB: Job posting is only available to Bioinformatics.ca members. To become a member please sign up for an account here. Job postings are moderated and must be approved before posting becomes public.

For full-time positions outside of Canada, please include the statement: "Applications from Canadian citizens are welcome".

Job Title	Institution/Company	Location	Date Posted
Junior Software Developer	Ontario Institute for Cancer Research	Toronto, On	2016-09-02 13:23
Hong Kong PhD Fellowship Scheme 2017/18	City University of Hong Kong	Hong Kong	2016-08-31 06:29
Post Doctoral Position in Tuberculosis	University of British columbia	Vancouver, BC.	2016-08-30 18:36
Postdoctoral fellowship in Computational Biology	University of British Columbia	Vancouver, BC, Canada	2016-08-30 08:08
Post-doctoral Fellow	Ontario Institute for Cancer Research	Toronto, ON	2016-08-29 13:48
Postdoctoral Researcher - Community Genome Database Development	University of Tennessee, Knoxville	Knoxville, TN, United States	2016-08-26 14:08
Research Associate	University of Saskatchewan	Saskatoon, SK	2016-08-23 16:40
Bioinformatics specialist and Web Database Developer	McGill Centre for Integrative Neuroscience	Montreal, QC	2016-08-16 16:02
Postdoctoral Fellowship in Bioinformatics	McGill University	Montreal	2016-08-16 12:22
Post-Doctoral fellow in High Dimensional Flow Cytometry Bioinformatics	Simon Fraser University / University of British Columbia	Vancouver, BC	2016-08-15 01:47

The global bioinformatics market, **valued at nearly $3.2 billion in 2012**, is forecast to grow to nearly $7.5 billion by 2017, according to Wellesley, Mass.-based BCC Research.

Healthcare IT News, April 30, 2013
Bioinformatics grows by billions by Bernie Monegain
`http://www.healthcareitnews.com/news/bioinformatics-grows-billions`

Market Size

- Bioinformatics Market worth **16.18 Billion** USD by 2021
  - `https://www.marketsandmarkets.com/PressReleases/bioinformatics-market.asp`
- Bioinformatics Market Size Worth **US$ 16 Billion** By 2022
  - `https://www.marketwatch.com/press-release/bioinformatics-market-size-worth-us-16-billion-by-`

What/**Who** is a bioinformatician?

According to a (dated?) survey on www.bioinformatics.org (540)

- Biology (192) 36%
- Computer Science (133) 25%
- Engineering (72) 13%
- Mathematics (26) 5%
- Physics (20) 4%
- Chemistry (34) 6%
- Other (54) 10%

Professional **associations**

ISCB — International Society for Computational Biology (www.**iscb**.org)
SMB — Society for Mathematical Biology (www.**smb**.org)
CSSB — Canadian Society for Systems Biology (www.**sysbiosociety**.ca)

Essential Cellular Biology: Molecules

- Deoxyribonucleic acid (**DNA**)
- Ribonucleic acid (**RNA**)
- **Proteins**

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**
- **Code of instructions** for life

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**
- **Code of instructions** for life
  - A **list of parts** and a **user manual**

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**
- **Code of instructions** for life
  - A **list of parts** and a **user manual**
- Each one of your cell has an **identical** copy of your **DNA**

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

- DNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **T**
- **Code of instructions** for life
  - A **list of parts** and a **user manual**
- Each one of your cell has an **identical** copy of your **DNA**
- **Different regions of your DNA are active** in each cell

Essential Cellular Biology: Deoxyribonucleic acid (DNA)

Species	**Size**
Potato spindle tuber viroid (PSTVd)	360
Human immunodeficiency virus (HIV)	9,700
Bacteriophage lambda (λ)	48,500
Mycoplasma genitalium (bacterium)	580,000
Escherichia coli (bacterium)	4,600,000
Drosophila melanogaster (fruit fly)	120,000,000
Homo sapiens (**human**)	3,000 000,000
Lilium longiflorum (easter lily)	90,000,000,000
Amoeba dubia (amoeba)	670,000,000,000

Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**

Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **U**

Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **U**
- **Tens**, **hundreds**, **thousands** nucleotides (letter) long

Essential Cellular Biology: Ribonucleic acid (RNA)

- RNA is a **polymer**
- Can be seen as a **string** over four letters: **A**, **C**, **G**, **U**
- **Tens**, **hundreds**, **thousands** nucleotides (letter) long
- Gene **transcription** and **translation**, but also **regulation**, **editing**, etc.

Essential Cellular Biology: Proteins

- Protein is a **polymer**

Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: **A**, **C**, **D**,...**Y**

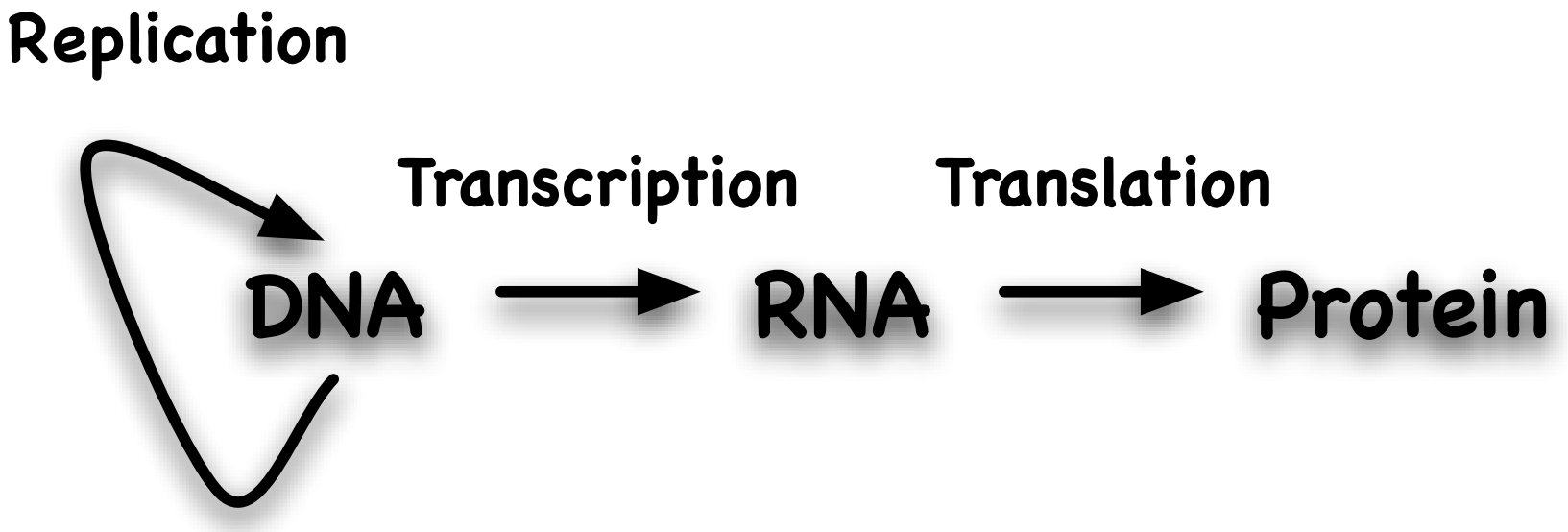
Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: **A**, **C**, **D**,...**Y**
- **Hundreds** or **thousands** amino acids (letter) long

Essential Cellular Biology: Proteins

- Protein is a **polymer**
- Can be seen as a **string** over twenty letters: **A**, **C**, **D**,...**Y**
- **Hundreds** or **thousands** amino acids (letter) long
- **Catalytic** activity, **transporter** activity, **binding**, etc.

Essential Cellular Biology: Central Dogma



Essential Cellular Biology: Gene Definition

What is a gene?

- A locatable **region** of genomic sequence [DNA], corresponding to a unit of inheritance, which is associated with **regulatory** regions, **transcribed** regions, and or other **functional** sequence regions.

Essential Cellular Biology: Gene Definition

What is a gene?

- A locatable **region** of genomic sequence [DNA], corresponding to a unit of inheritance, which is associated with **regulatory** regions, **transcribed** regions, and or other **functional** sequence regions.

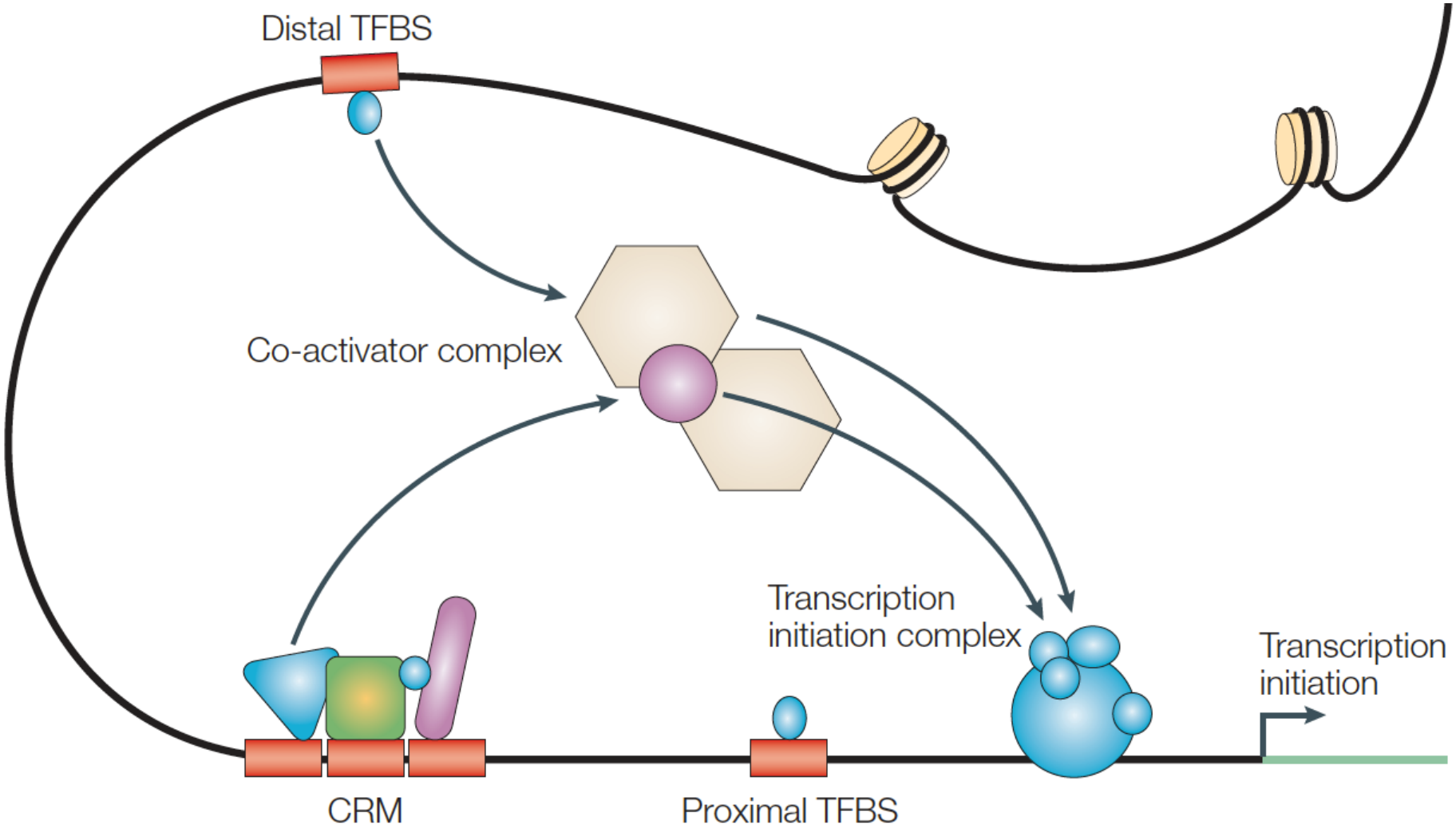
Pearson H.
Genetics: what is a gene?.
Nature 441 (7092): 398–401, 2006.

Essential Cellular Biology: Gene Regulation

Gene Regulation

- Ensemble of mechanisms by which the cell **increases** or **decreases** the production of (protein or RNA) **genes**

Essential Cellular Biology: Gene Regulation



Wasserman & Sandelin (2004) Nature Reviews Genetics 5(4), 276–287.

Essential Cellular Biology: Information

- Molecular **sequences** and **structures**
- Hundreds of **ontologies** describe the parts and processes
- **High-throughput experiments**
  - **ChIP-Seq** informs about protein-DNA interactions
  - **DNA microarrays** measure the **expression** levels of genes
  - And many more

Essential Cellular Biology: Resources

!xe unlockinglifescode.org

See also:

- https://www.ted.com/talks/james_watson_on_how_he_discovered_dna.

References


R. Durbin, S. Eddy, A. Krogh, and G. Mitchison.
Biological Sequence Analysis.
Cambridge University Press, 1998.


D. Gusfield.
Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.
Cambridge University Press, 1997.


N. C. Jones and P. A. Pevzner.
An introduction to bioinformatics algorithm.
MIT Press, 2004.

Pensez-y!

L'impression de ces notes n'est probablement pas nécessaire!